

Anonymous Gossiping

Anwitaman Datta
NTU Singapore

Abstract

In this paper we introduce a novel gossiping primitive to support privacy preserving data analytics (PPDA). In contrast to existing computational PPDA primitives such as secure multiparty computation and data randomization based approaches, the proposed primitive “anonymous gossiping” is a communication primitive for privacy preserving personalized information aggregation complementing such traditional computational analytics. We realize this novel primitive by composing existing gossiping mechanisms for peer sampling & information aggregation and onion routing technique for establishing anonymous communication. This is more an ‘ideas’ paper, rather than providing concrete and quantified results.

Keywords: privacy, anonymity, aggregation, gossip algorithms

“It is perfectly monstrous the way people go about nowadays saying things against one, behind one’s back, that are absolutely and entirely true.” — Oscar Wilde

1 Introduction

Information aggregation and mining is often used to obtain collective intelligence and generate a panoramic (macroscopic) view of a system or to devise useful recommendation mechanisms. An interesting niche which has been studied for the last decade is that of privacy preserving data mining (PPDM). The essential idea, to quote the seminal paper on PPDM [1], is “*Since the primary task in data mining is the development of models about aggregated data, can we develop accurate models without access to precise information in individual data records?*”. The early works on PPDM were based on random perturbation of information. Since then, a new class of PPDM based on secure multiparty computation [14] has also evolved. The trade-offs between the two approaches are mainly on accuracy and computational complexity & scalability. Research on both these individual families of privacy preserving data analytics (PPDA) as well as hybridized solutions continue in full steam. Privacy preserving data mining in P2P environ-

ments [4] has also gained considerable attention in recent years.

In this paper we address an orthogonal question. *Can we facilitate collaborative data analytics among users without disclosing the identity of who are participating and contributing the data?* Computational PPDA’s do not provide anonymity. Such privacy is important, say, when the analytics is carried out for a specific subset of users with some shared characteristics, such that besides the privacy of the individual records, the users may be interested to even preserve privacy in terms of them having those characteristics.

We propose a communication primitive (anonymous gossiping¹) which facilitates such privacy. Specifically, we adapt a well studied point-to-point anonymized communication technique, onion routing [10, 8] to achieve it. Note that the actual data analytics itself however may be additionally with or without preserving privacy of the individual records. For example, for an anonymized paper submission, it is ok that the reviewers can read the content of the paper, as long as they do not know who wrote the paper. It is in this sense that our mechanism compliments the existing computational primitives. Anonymous gossiping finds ready usage in emerging P2P applications such as user affinity based personalized decentralized search [13, 3] & personalized recommendation in decentralized online social networks [5].

While anonymous communication in p2p systems is an old and well studied problem, e.g., Freenet [7], the novelties of this paper are (i) defining a novel communication primitive for PPDA, and (ii) proposing a concrete way to do so by composing well studied existing building blocks.

In this short paper, we limit ourselves to defining this new problem and sketching a first solution for the same. A more rigorous analysis and evaluation of the proposed mechanism’s security, performance, overheads and subsequent necessary optimization or exploration of alternatives are all issues for future study.

In Section 2 we provide a more succinct description of the problem along with a sketch of the solution. We elabo-

¹Epidemic information dissemination leveraging anonymous interactions in mobile ad hoc network has been studied in the past, and uses the same name [6], but what we do is completely unrelated.

rate in detail our assumptions and notations in Section 3 before providing the anonymous gossiping protocol in Section 4. We wrap up in Section 5 with concluding remarks highlighting several interesting extensions and research problems that present themselves from the current work.

2 Problem statement & solution sketch

We want to facilitate the following:

1. Allow user specific information (lets call it *user profile*) to be used to carry out any kind of personalized aggregation/clustering, etc. Such mechanisms can then be used for various personalized services such as recommendation or query expansion [3, 13].
2. Ensure that an user can not be associated with her² profile by others, even while individual users benefit from personalization facilitated by analysis of information aggregated from other similar users.

The basic *outline* of a potential solution to achieve the above mentioned objectives comprise of the following steps and building blocks:

1. Aggregation task delegation: Each user delegates the task of personalized aggregation to a proxy peer.
2. Proxy peers interact among each other to carry out the aggregation task on behalf of the users.
3. Ensure that the proxy peer is oblivious of the identity of the user(s) on whose behalf it carries out aggregation task. This in turn will ensure users' privacy. This necessitates a mechanism for users to assign the task to a proxy without being identified, and also a mechanism for the proxy to still be able to deliver back the aggregated information to the original user without knowing who it is.

Here we describe a mechanism (anonymous gossiping) to achieve the last point. How the aggregation task itself is carried out among the proxies is an orthogonal issue. This includes the issues of both how proxies interact among each other, and how they carry out the data analytics. Anonymous gossiping is generic in that it can be applied to provide user privacy while using arbitrary gossiping algorithms for information aggregation.

Whether any peer is adequate to act as a proxy, or whether some other considerations such as trustworthiness, or betweenness in social graph (facilitating quicker aggregation) etc. need to be taken into account while delegating the task is ignored in the current work.

²For simplicity, we choose to use the feminine form to address the users, instead of using his/her/its on every occasion.

3 Assumptions and notations

Our solution relies on the following assumptions and existing primitives.

1. Users form and participate in an overlay. This overlay may be a classical unstructured network or a semantic or social overlay.
2. Users use public key as their logical identifier in the system. However, there is no need for a public key infrastructure (PKI) since we are not trying to establish if a specific public key belongs to a specific user or not. Public key is used so that anything signed with it can be decrypted by only its corresponding private key.
3. A random set of peers (public keys and corresponding contact information such as IP address/port number) can be obtained without an adversary knowing who obtained a specific information. This assumption is important, otherwise, if one can determine who all obtained a specific peer ID from the sampling service, then it reduces the degree of anonymity.

A random set of peers can readily be obtained using gossip based peer sampling [12]. We argue that peers who participate in the process of peer sampling for a relatively long time would encounter sufficiently large number of other peers to mitigate any set intersection analysis by an adversary.

4. Proxies delegated by users of similar profiles can discover each other and carry out the aggregation task. Note that this last assumption is needed for the aggregation task, and is orthogonal to the anonymity issues. Gossip based mechanism like T-man [11] or variants [16] can be applied for this. Note also that the gossiping overheads for the various tasks like peer sampling and information aggregation can be amortized.

We use the following notations while detailing the anonymous gossiping mechanism.

-
- α_i Public key and ID of peer i .
 - $E_i(.)$ Encryption of message with public key of peer α_i , so that only she can decrypt it.
 - Φ_i Random set of peers that peer α_i has obtained somehow.
 - κ_i A symmetric en/de-cryption key created by peer α_i .
 - $\kappa_i(.)$ Message encrypted with the key κ_i .
 - π_i Profile of peer α_i . The records of the profile may/not need themselves to be perturbed or obfuscated for privacy preserving data analytics [1]. That is however an orthogonal issue.
 - π'_i Aggregated/personalized clustered information corresponding to profile π_i .
-

4 Anonymous gossiping

There are three logical phases for anonymous gossiping: (phase I) aggregation task delegation to a proxy in an anonymous manner, (phase II) the proxies carrying out the delegated aggregation tasks, and finally (phase III) obtaining the results back from the delegate in an anonymous manner.

The aggregation task (phase II) is an interesting problem on its own right. Either existing solutions [9, 2, 13] or new ones may be applied for it. We consider it as a black-box and focus on the other two phases as described next.

In the description below we consider a scenario where *Alice* is using anonymous gossiping to carry out privacy preserved personalized aggregation.

4.1 Delegation of aggregation task

Aggregation task is delegated to a proxy as follows:

- Obtain Φ_{Alice} , a moderately large and random subset of peers in the system, e.g., using peer-sampling [12].
- Determine the candidate α_μ for task delegation where $\alpha_\mu \in \Phi_{Alice}$.
Alice needs to send the message $msg = (\pi_{Alice}, \kappa_{Alice})$ containing her profile and a symmetric key to this delegate anonymously.
 Note that the message contains the profile, but not *Alice*'s identity. However, if the profile itself contains identity revealing details (such as search terms from 'ego search') then our approach can not provide anonymity. Generally speaking, possible obfuscation of the records using traditional PPDM techniques [1] may additionally be necessary.
- Choose k other peers $\alpha_i \in \Phi_{Alice}$, where k is a random integer chosen uniformly from some predefined range, say [5, ..., 20]. These peers will be used to form an onion route [8, 10] between *Alice* and the delegate α_μ .
- Send to peer α_{ρ_1} an onion encoded message $E_{\rho_1}(E_{\rho_2}(\dots(E_{\rho_{k-1}}(E_{\rho_k}(E_\mu(msg), \mu), \alpha_\mu), \alpha_{\rho_k}), \dots), \alpha_{\rho_2}))$

When a peer α_{ρ_j} receives an onion encoded message from $\alpha_{\rho_{j-1}}$, she can only decrypt the outermost layer with her own private key. Upon decryption, it obtains another encrypted message along with the identity/address of the next node to which it should pass the same. Thus, intermediate nodes do not know how many nodes or who have already routed the message, nor the nodes who will subsequently do so. Each node only knows its immediate up & down-stream peers for an onion route. Given that the nodes were chosen at

random by the source further reduces chances of collusion. The random choice of the length k of the onion route provides a further level of obfuscation and plausible deniability for *Alice*. It also provides robustness against small scale opportunistic collusion among some of the intermediate nodes to unravel the route requester's identity. Onion routing is robust against traffic snooping provided a minimal amount of ambient traffic is present in the system [15].

Once the designated delegate α_μ obtains the $msg = (\pi_{Alice}, \kappa_{Alice})$, it can carry out the Phase-II task of aggregation and analytics using π_{Alice} to compile π'_{Alice} .

4.2 Collecting aggregated information

One option to collect the aggregated information is for *Alice* to probe the delegate, and pull the response along a (new) onion route. Note that *Alice* should not reveal her identity to the delegate, so the delegate can not know which node to send the response to. The response $\kappa_{Alice}(\pi'_{Alice})$ encrypted with the symmetric key originally sent by *Alice*, and digitally signed by the delegate α_μ may be sent upstream along the onion route (since each node knows the immediate up & down-stream neighbors of a route without the delegate having to know specifically the destination).

There are however some potential drawbacks with such a pull based approach. Firstly, since *Alice* does not know if and when the delegate has completed computation of π'_{Alice} , she may have to initiate pull on multiple occasions. Secondly, and more fundamentally, a passive attacker can monitor the network traffic (sniffing for $\kappa_{Alice}(\pi'_{Alice})$) and detect the terminal point. This can be alleviated if the message is encrypted by the intermediate nodes at every hop using the public key of the immediate node upstream.

An alternative to pull is a blind gossip based push mechanism. Whenever proxy α_μ needs to send the aggregated information π'_{Alice} back to *Alice*, she can just flood the network with corresponding $\pi'_{Alice}, \alpha_\mu, \kappa_{Alice}(\pi_{Alice})$ digitally signed with its private key. On receiving any such flooded message originating from α_μ , *Alice* can still determine whether the message is indeed meant for her or not using $\kappa_{Alice}(\pi_{Alice})$ even if α_μ is also proxy for other peers. *Alice* should continue forwarding the message in all cases, so that no one monitoring the network can identify her as the intended destination for the information. A possible optimization during flooding is that each peer propagates the message to only peers it has obtained onion routed messages within a past time window. That will ensure, in absence of churn, that the source (*Alice*) gets the aggregated information, while avoiding a larger scale flooding.

5 Concluding remarks

We have defined a new communication primitive, namely *anonymous gossiping*, which can support privacy preserving data aggregation and analytics complementing traditional computational primitives for PPDA based on randomization or secure multi-party computation. We have provided a rough but concrete sketch of one way to realize anonymous gossiping by composing existing techniques like peer sampling and onion routing. Use of such mature techniques is expected to facilitate a quick implementation of anonymous gossiping.

Additionally, this paper opens interesting avenues spanning algorithms, implementation as well as analysis which forms our ongoing work. Exploration of new, more efficient, robust and churn resilient algorithms for anonymous gossiping is one direction. Clever implementation, particularly amortizing the various gossiping overheads (needed during peer-sampling and aggregation) provide nice systems design opportunities. Threat analysis including quantifying the trade-offs between the degree of anonymity and the time and messaging overheads in the peer-sampling process is a third frontier.

Acknowledgements

This work was inspired from discussions with Anne-Marie Kermarrec in the context of the GOSSPLE project during a research visit by the author to the ASAP Inria research group pertaining to the said project.

References

- [1] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. *ACM SIGMOD Records*, 29(2), 2000.
- [2] Farnoush Banaei-Kashani and Cyrus Shahabi. Swam: A family of access methods for similarity-search in peer-to-peer data networks. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004.
- [3] Matthias Bender, Tom Crecelius, Mouna Kacimi, Sebastian Michel, Thomas Neumann, Josiane Xavier Parreira, Ralf Schenkel, and Gerhard Weikum. Exploiting social relations for query expansion and result ranking. In *ICDE Workshops*, 2008.
- [4] Kanishka Bhaduri, Kamalika Das, and Hillol Kargupta. Peer-to-peer data mining, privacy issues, and games. In *Autonomous Intelligent Systems: Multi-Agents and Data Mining Workshop (AIS-ADM)*, 2007.
- [5] Sonja Buchegger and Anwitaman Datta. A case for P2P infrastructure for social networks - opportunities and challenges. In *Proceedings of WONS 2009, The Sixth International Conference on Wireless On-demand Network Systems and Services*, 2009.
- [6] Ranveer Chandra, Venugopalan Ramasubramanian, and Kenneth P. Birman. Anonymous gossip: Improving multicast reliability in mobile ad-hoc networks. *Distributed Computing Systems, International Conference on*, 2001.
- [7] Ian Clarke, Oskar Sandberg, Brandon Wiley, and Theodore W. Hong. Freenet: A distributed anonymous information storage and retrieval system. *Lecture Notes in Computer Science*, 2001.
- [8] Roger Dingledine, Nick Mathewson, and Paul Syverson. Tor: The second-generation onion router. In *SSYM'04: Proceedings of the 13th conference on USENIX Security Symposium*, pages 21–21, Berkeley, CA, USA, 2004. USENIX Association.
- [9] Hector Garcia-Molina and Arturo Crespo. Semantic overlay networks for p2p systems. Technical Report 2003-75, 2003.
- [10] David Goldschlag, Michael Reed, and Paul Syverson. Onion routing. *Commun. ACM*, 42(2), 1999.
- [11] Mark Jelasity and Ozalp Babaoglu. T-man: Gossip-based overlay topology management. In *The Fourth International Workshop on Engineering Self-Organizing Applications (ESOA'06)*, 2006.
- [12] Márk Jelasity, Spyros Voulgaris, Rachid Guerraoui, Anne-Marie Kermarrec, and Maarten van Steen. Gossip-based peer sampling. *ACM Trans. Comput. Syst.*, 25(3), 2007.
- [13] Anne-Marie Kermarrec. Challenges in Personalizing and Decentralizing the Web: An Overview of GOSSPLE. In *SSS '09: Proceedings of the 11th International Symposium on Stabilization, Safety, and Security of Distributed Systems*, 2009.
- [14] Y. Lindell and B. Pinkas. Secure multiparty computation for privacy-preserving data mining. In *the Journal of Privacy and Confidentiality*, 1(1), 2009.
- [15] Andrei Serjantov and Peter Sewell. Passive-attack analysis for connection-based anonymity systems. *International Journal of Information Security*, 42(3), 2005.
- [16] Spyros Voulgaris and Maarten van Steen. Epidemic-style management of semantic overlays for content-based searching. 2005.